NYSBC
NEW YORK STRUCTURAL BIOLOGY CENTER

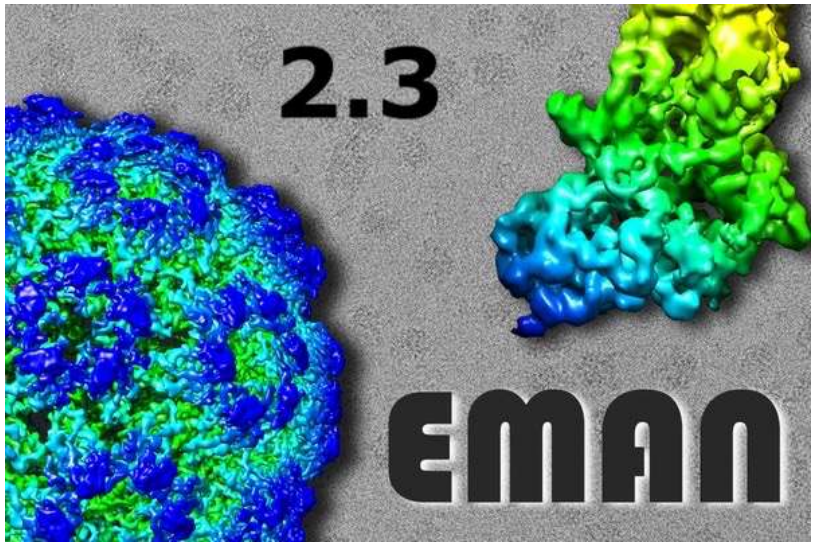
SIMONS ELECTRON MICROSCOPY CENTER

Winter-Spring 2025
EM Course

Data Processing
(Theory and algorithms)

Reza Khayat (CCNY/CUNY)

Some complete software packages



RELION

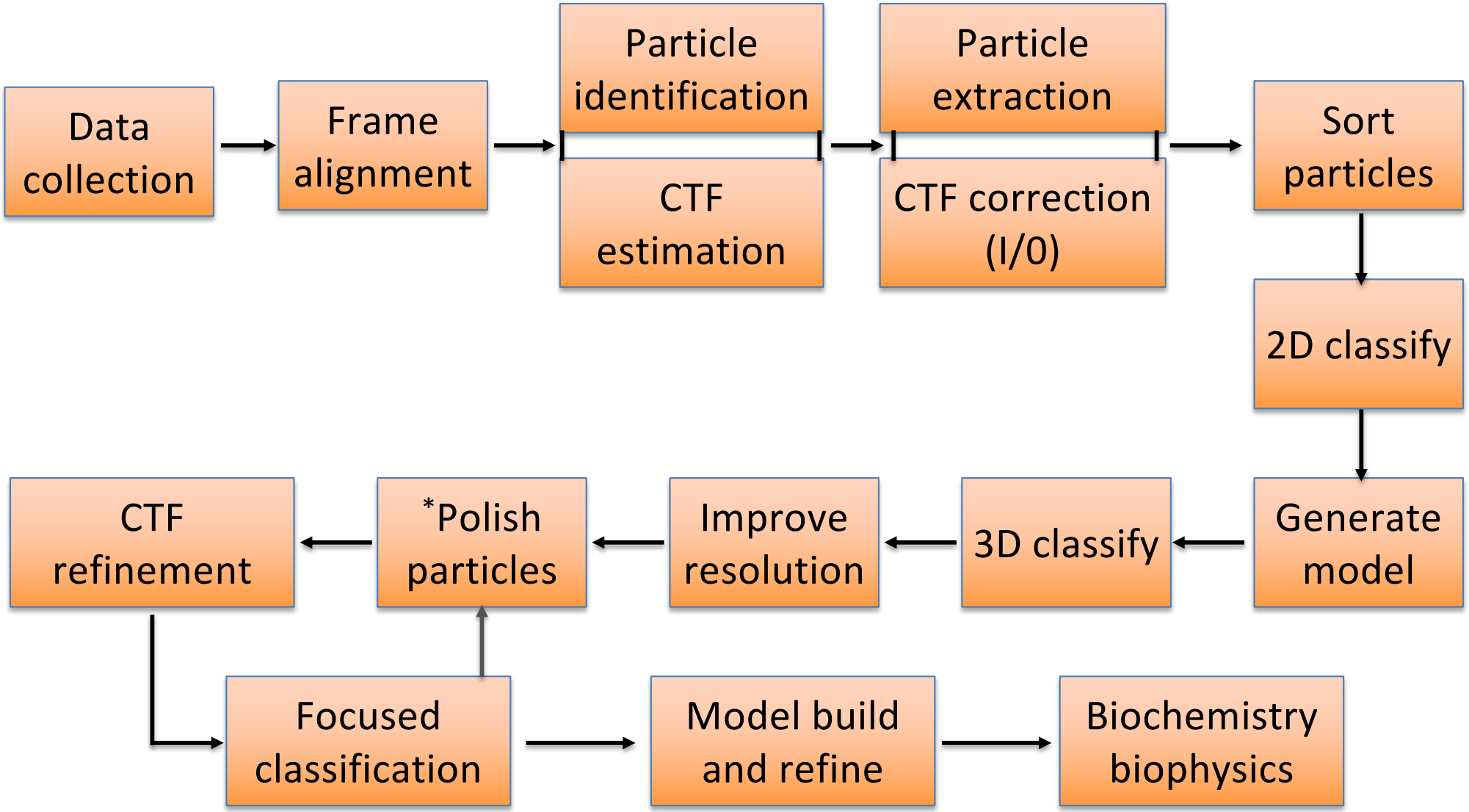
v3b1

CISTEM

Sphire/Sparx

cryoSPARC

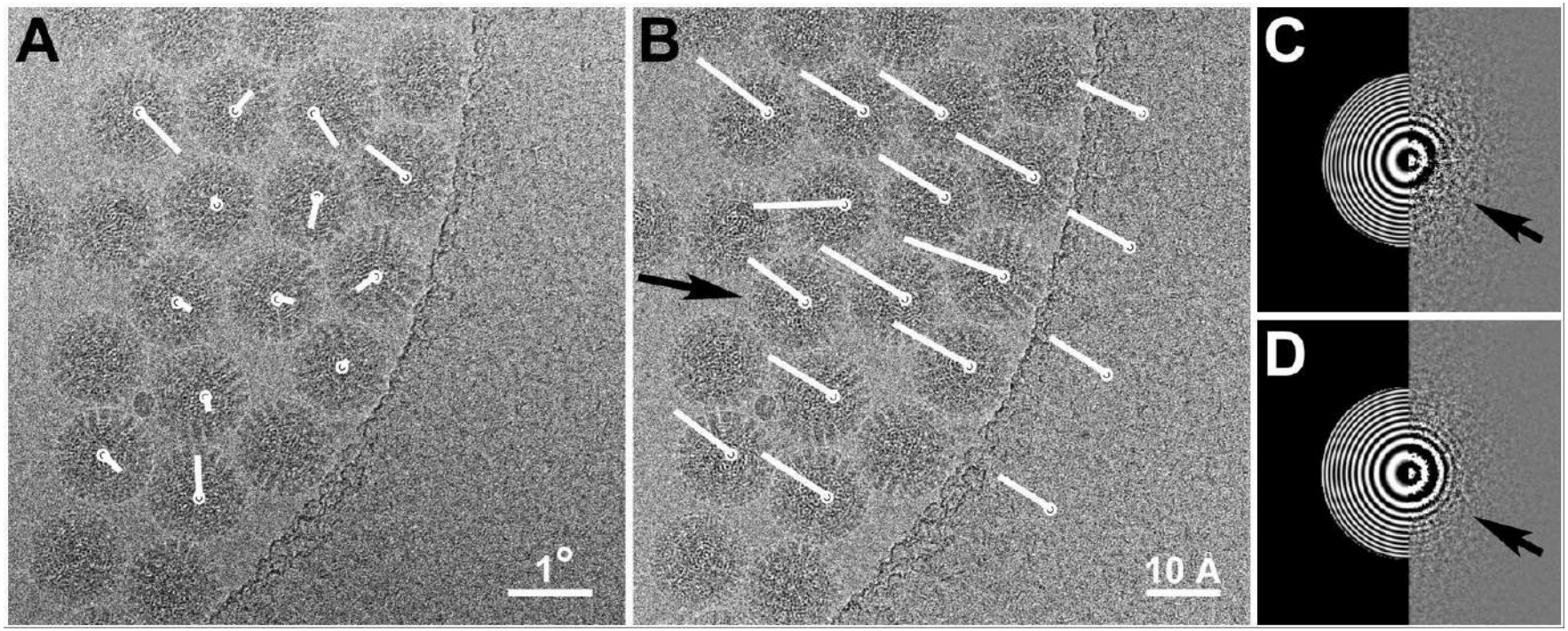
General SPA Image analysis



Workflow in cryoSPARC

Movie frame alignment

- UCSF MotionCor2
- Unblurr
- Warp
- cryoSPARC

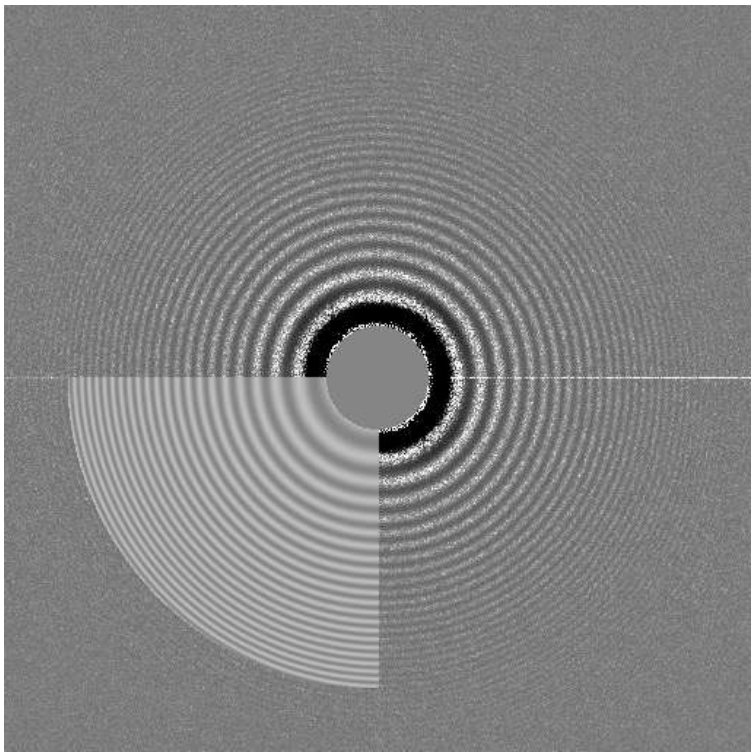


A: Unaligned micrographs with rotation vector
B: Unaligned micrographs with translation vector
C: Power spectrum of unaligned micrograph
D: Power spectrum of aligned micrograph

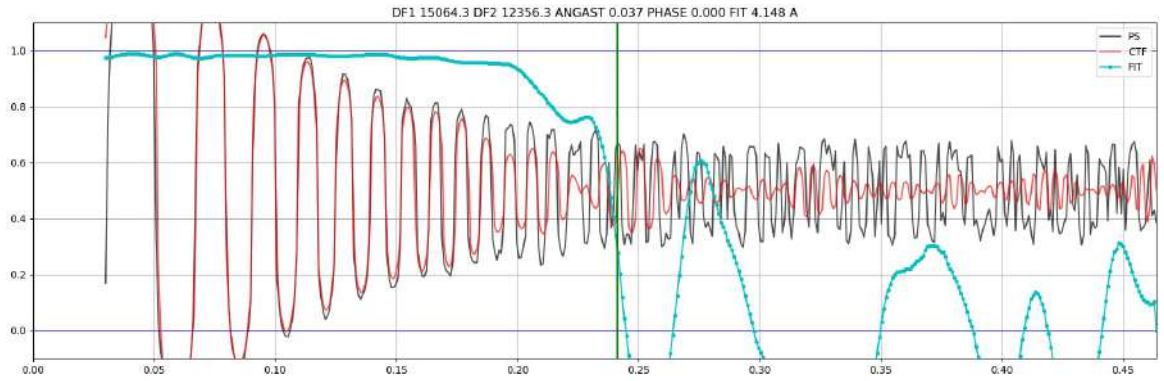
Campbell et al., 2012

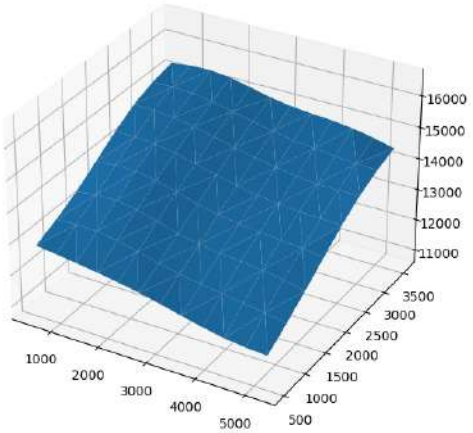
CTF Estimation

- CTFFIND4
- Sparx/EMAN
- GCTF
- Warp
- CryoSPARC (patch method)



CTFFIND4



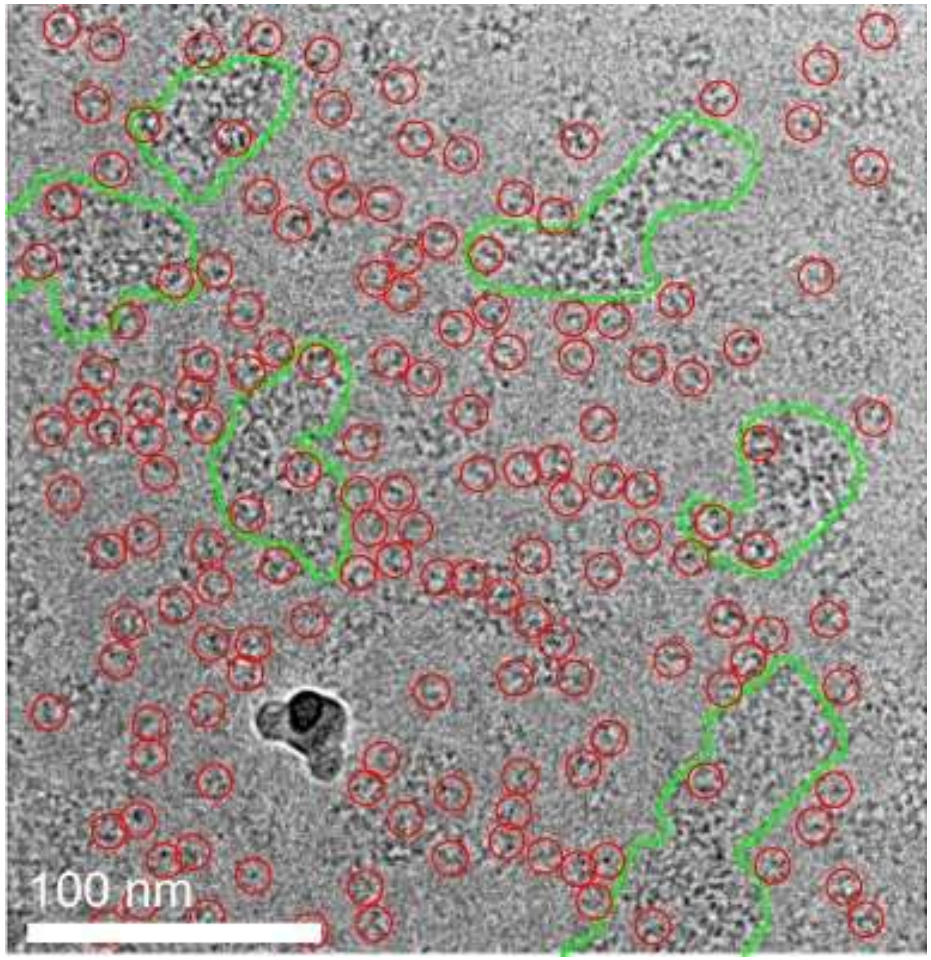


My stuff (CryoSPARC)

Particle Picking

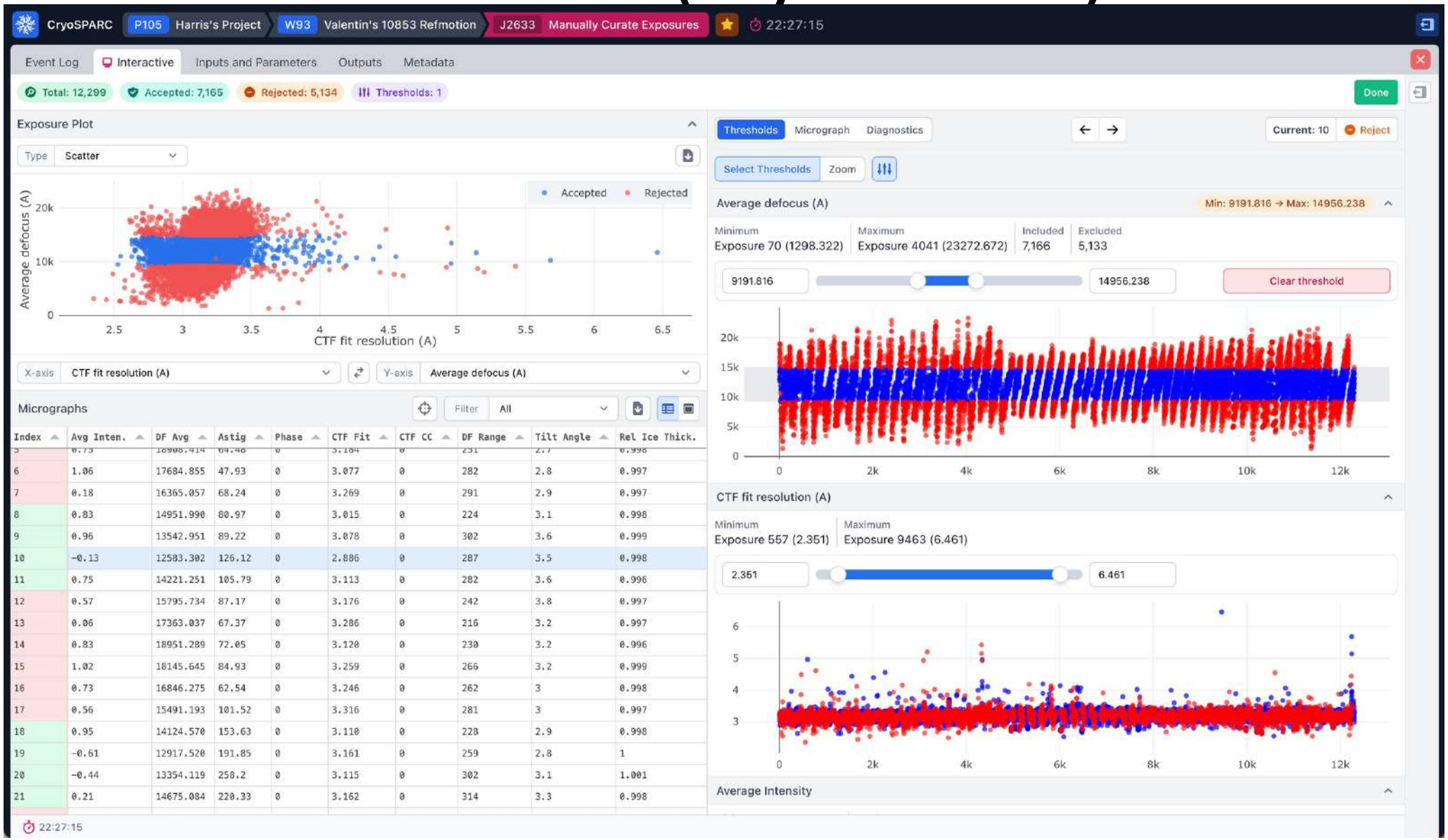
Manual – Blob – Template – Neural network

- Topaz
- Gautomatch
- Warp
- crYOLO
- FindEM
- Blob Picker
- DeepPicker
- DeepEM
- PIXER
- DRPnet
- DeepCryoPicker
- AutoCryoPicker
- *Start with provided model, get 2D classes, and retrain*



Bepler et al., 2019

Curation (cryoSPARC)



- Defocus range
- CTF fit resolution
- Number of particles

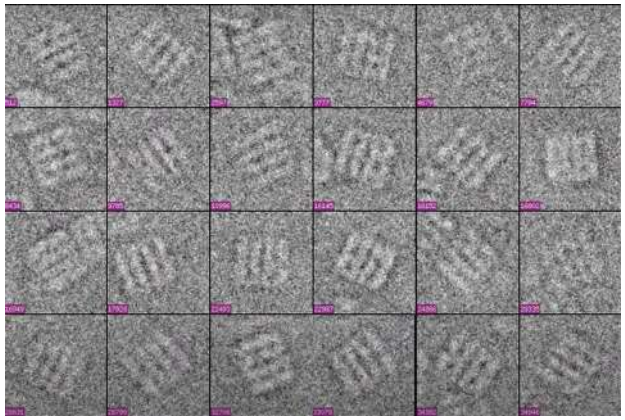
- Tilt angle
- Astigmatism
- Ice thickness

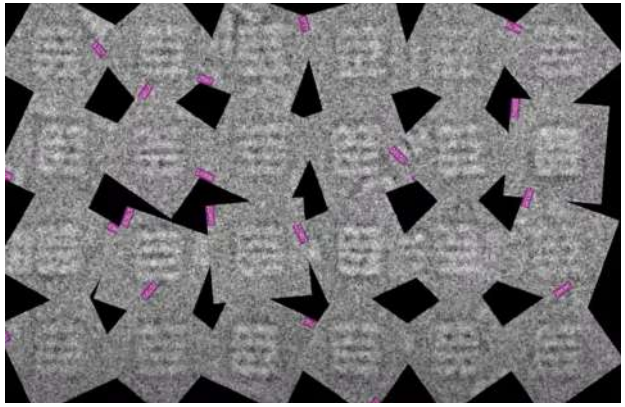
Particle Extraction

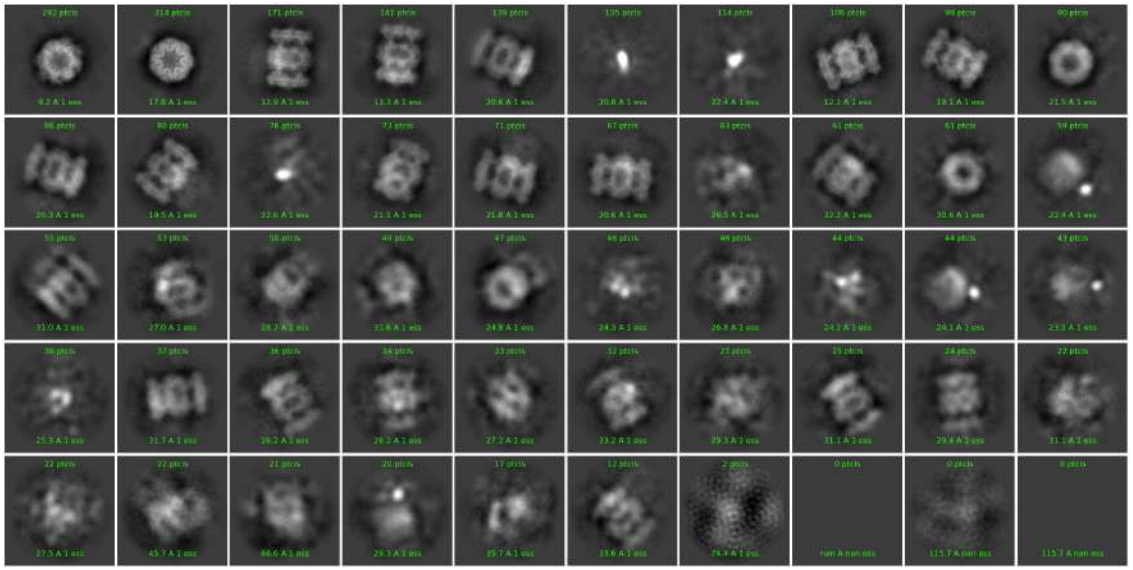
- Signal is delocalized according to energy of electron and defocus value of image:
  - Fred Sigworth 2022 lecture: r=delta*lambda*f
    - r is the radius that surrounds the particle. It describes how far the signal is delocalized.
    - delta=defocus in Å
    - lambda=0.02 Å (wavelength of e- at 300kV)
    - Frequency=desired resolution (e.g. 0.33$Å^{-1}$ for 3 Å)
- Even number with low prime factors (2, 3, 5, and 7)
  - I like 32, 64, 128, 256, 320, 384
- You may want to downsize (fourier bin) the particles to expedite initial data processing, and save on drive space. I downsize to 10Å/pixel or 64x64 (whichever first).

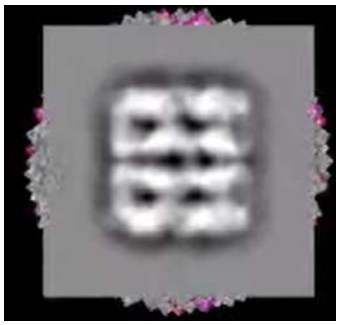
2D Classification

- cryoSPARC
- Relion
- Sparx/EMAN2
- ISAC
- Spider
- Simple
- *Remove “bad” particles*









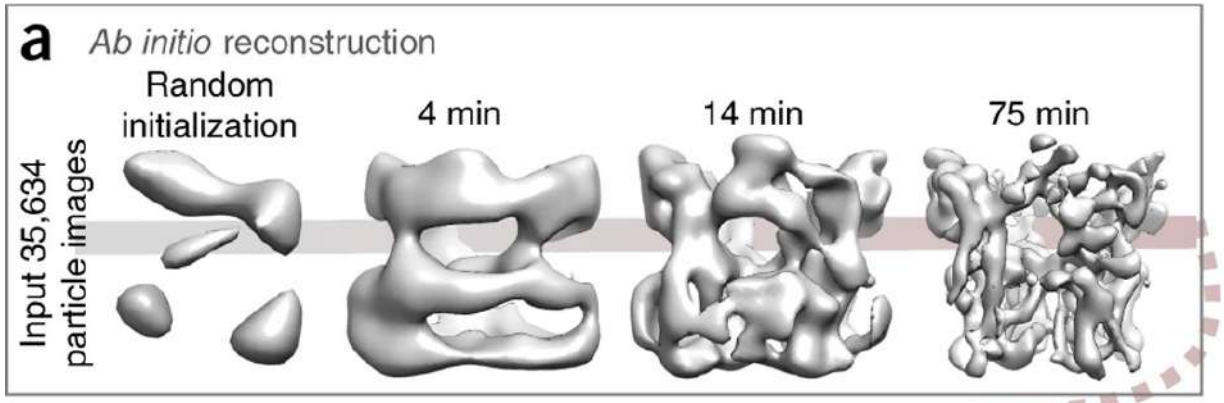
Lander 2009

2D Classification

- Anticipate 100 to 400 particles per class
- Don’t ask for too many classes
- I split my particle stack into stacks of 100K particles and process each separately to get clean-vs-dirty particles
  - Screen through various values for radius
  - Relion
    - Tau fudge
    - CTF
  - cryoSPARC
    - Turn off Force Max over poses/shifts
    - Initial classification uncertainty factory (2 and above)
    - Number of iteration to anneal sigma as high as 25
    - Set online-EM iterations to 40
    - Set Batchsize per class to 400
    - Change Re-center mask threshold (possibly as high as 0.75) for centering particles and smearing neighbors
    - set White noise model to off

Initial Model

- Random conical tilt
- Orthogonal conical tilt
- Common-lines
- Tomography with STA
- Random initial parameters, optimize with stochastic gradient descent (SIMPLE, cryoSPARC, and Relion).
- SAXS/SANS
- Structure prediction (calculate map of PDB)



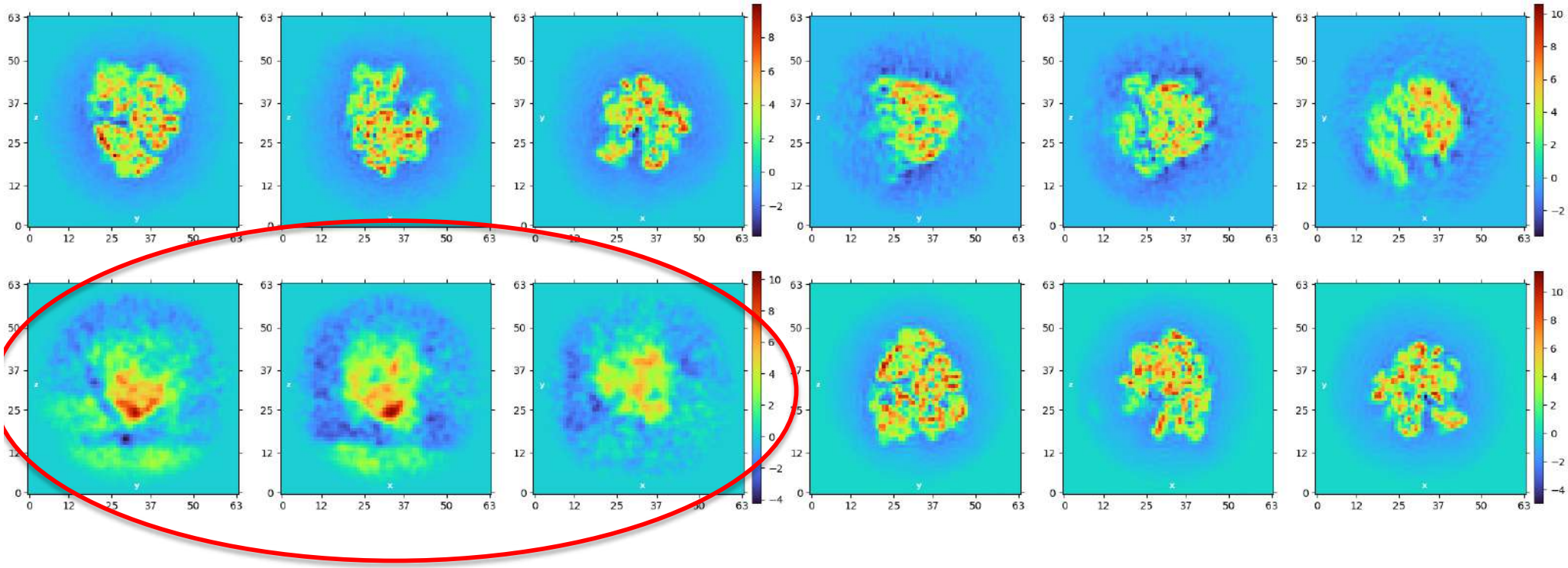
Punjani et al., 2017

Initial Model

- Generate multiple initial models if uncertain in model
  - Look for continuity in density
  - Look for sausages to indicate α-helices
  - Are projections comparable to class averages?
- Ask for multiple models to be generated
- CryoSPARC's starting frequency should have more information than particle_size / 5 (e.g. 300 / 5 = 60Å)
- Use C_1 symmetry

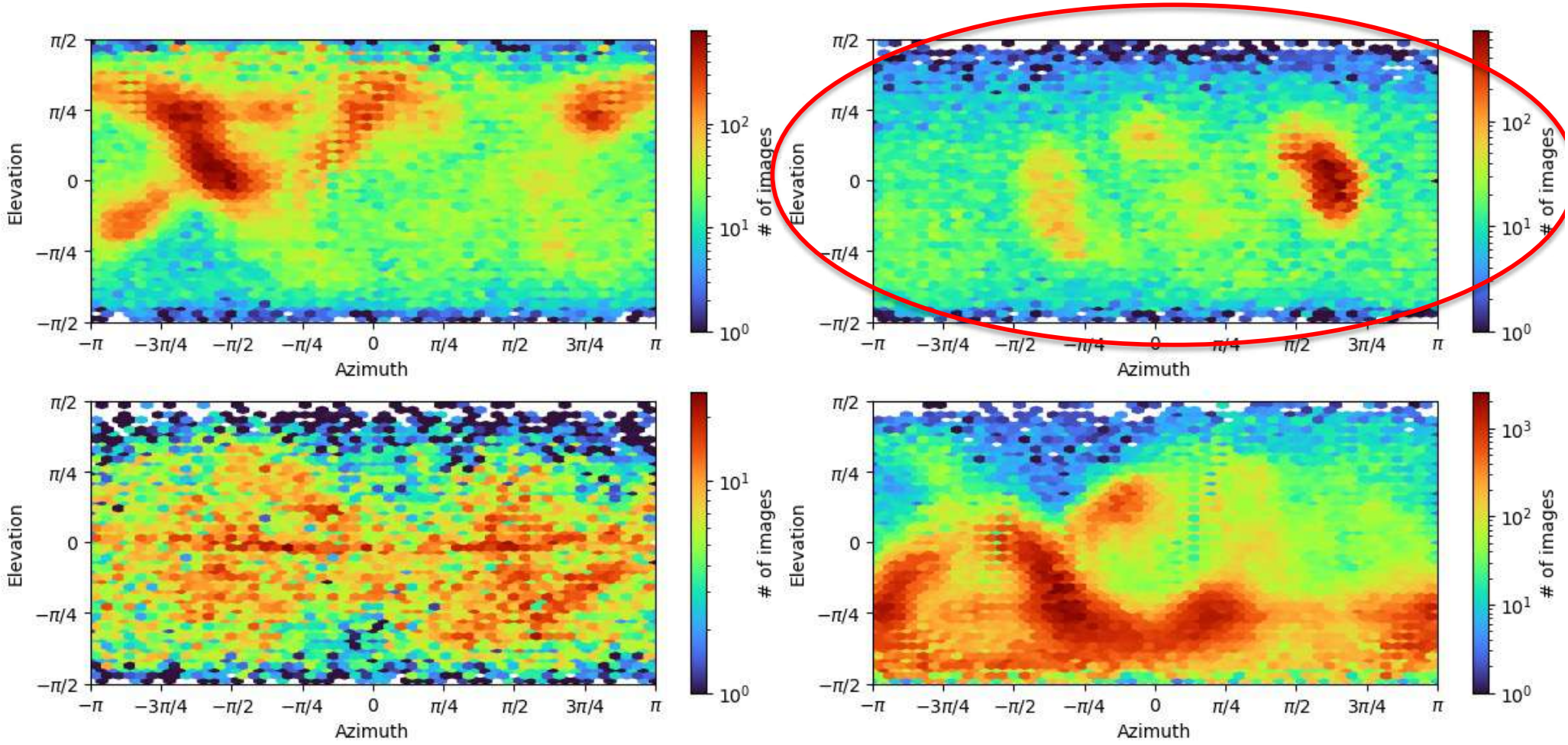
3D Classification

- Can be used to clean data further
  - Discard “bad” particles



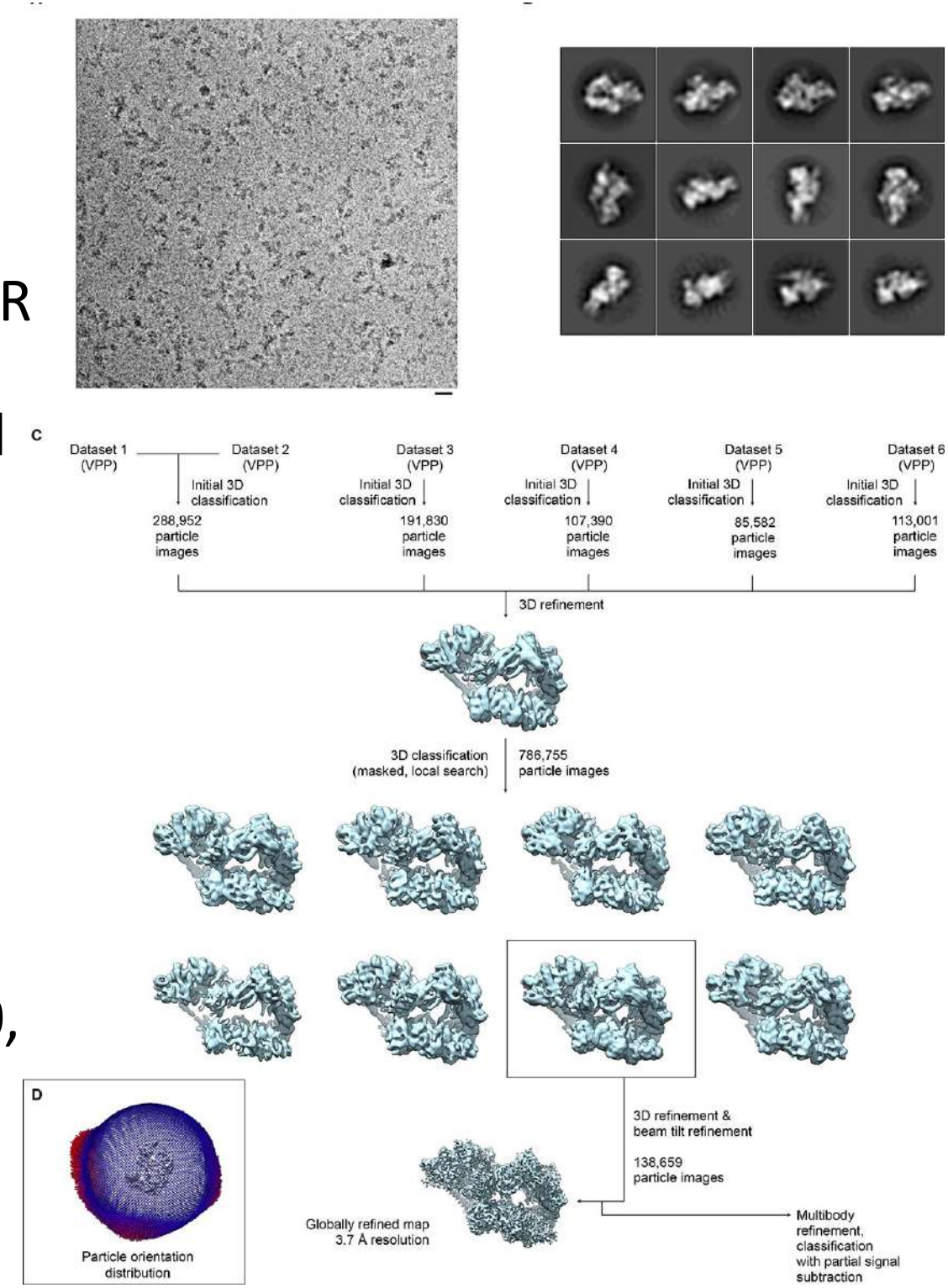
3D Classification

- Can be used to clean data further
  - Discard “bad” particles
  - Discard some preferred orientations



Enrich rare views

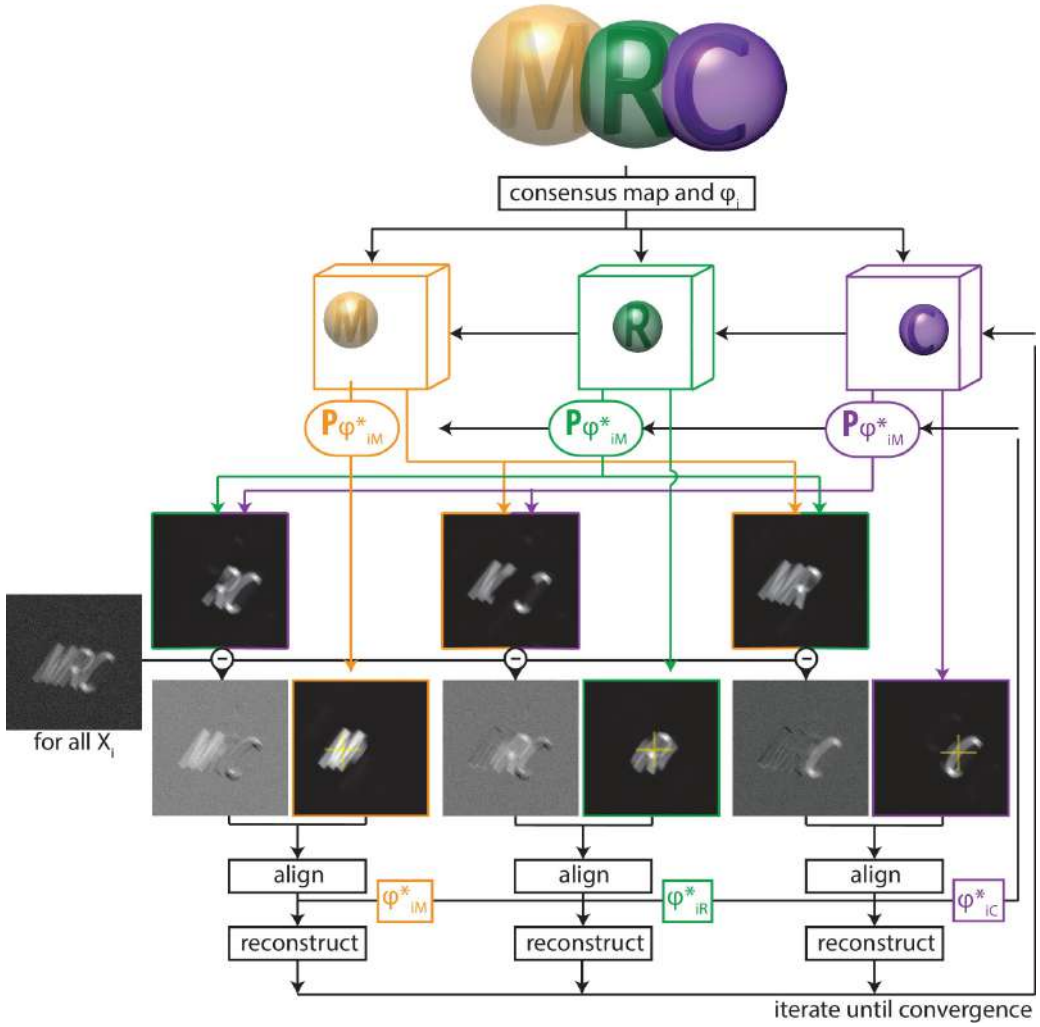
- Transcription Factor IIH (TFIIH), transcription initiation by RNAPII and NER
- Enrich rare views that 2D classification would discard
- Do 2D classification (B) for sanity check
- Extract particles and perform 3D classification with various tau2_fudge value sto enrich for rare views (D).
  - Value empirically determined. Try 1, 5, 10, 20, 50, 100, 200, 500, 1000



Greber et al., 2019

Improve local resolution

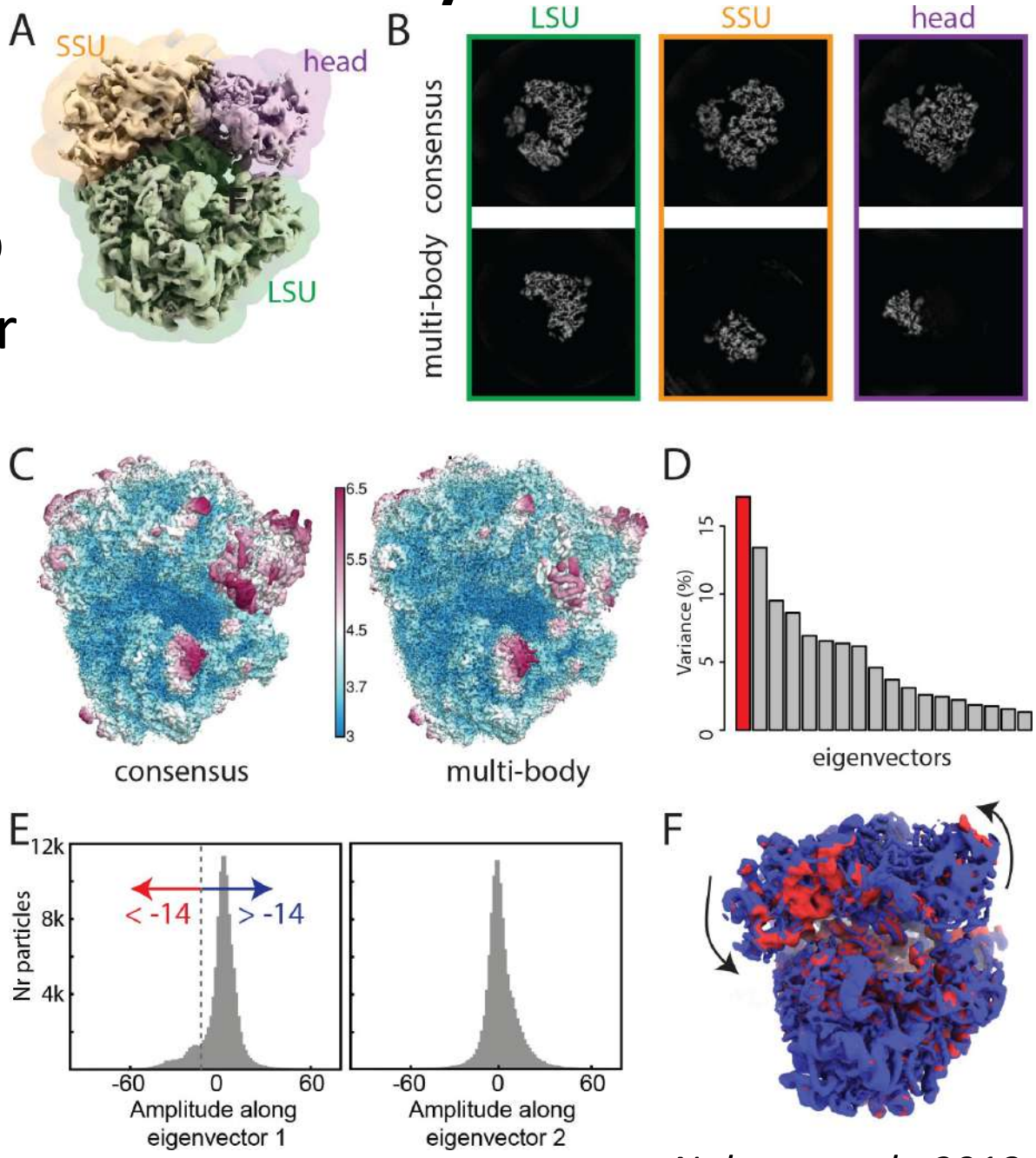
- Generate consensus map
- Domains/proteins moving independently
- Mask region of interest
- Use consensus map to subtract everything outside/inside of mask from each particle
- Use refinement parameters form consensus map to refine map of remaining signal
- Subtraction does not always work completely. May need to iterate through this process.



Nakane et al., 2018

Improve resolution identify motion

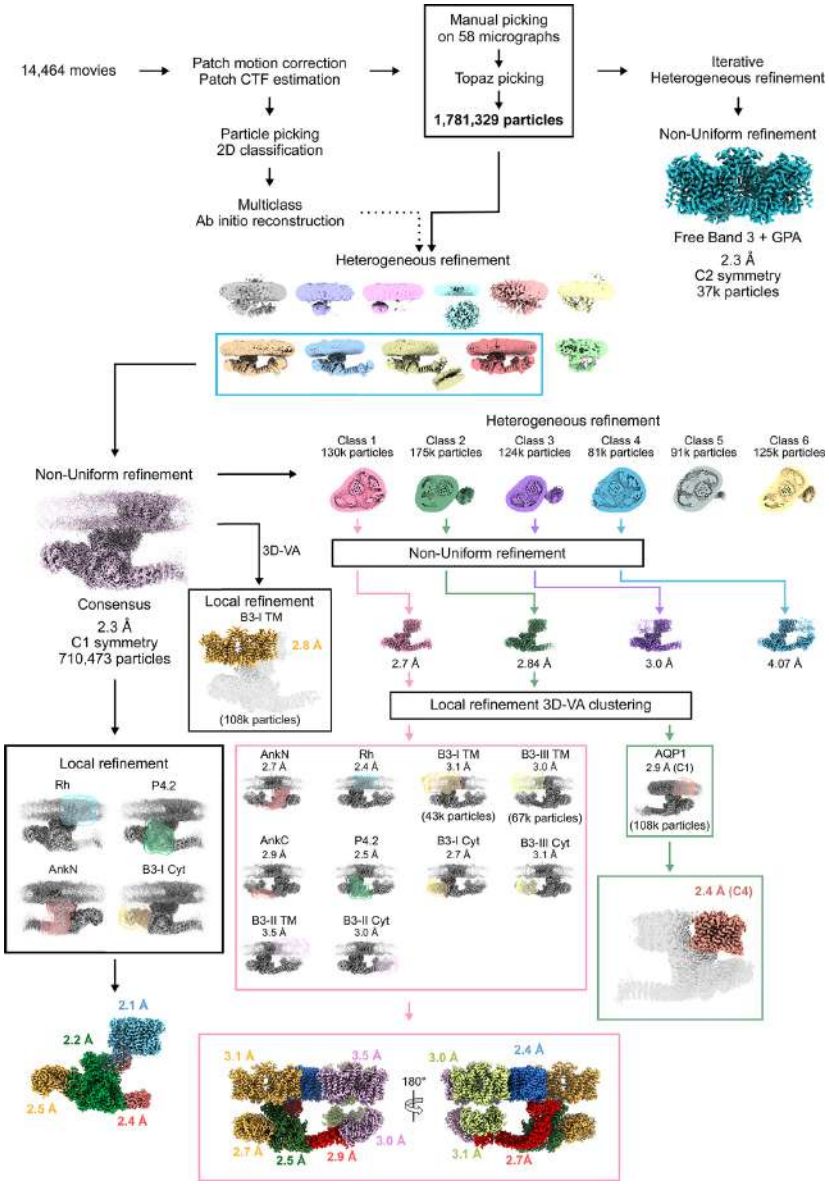
- Ribosome
- Generate consensus map
- Create body and mask for each region of interest
- Relion will do signal subtraction, local refinement, and PCA to identify rigid body motions
- Masked boundaries will not be trivial to interpret
- Relion



Nakane et al., 2018

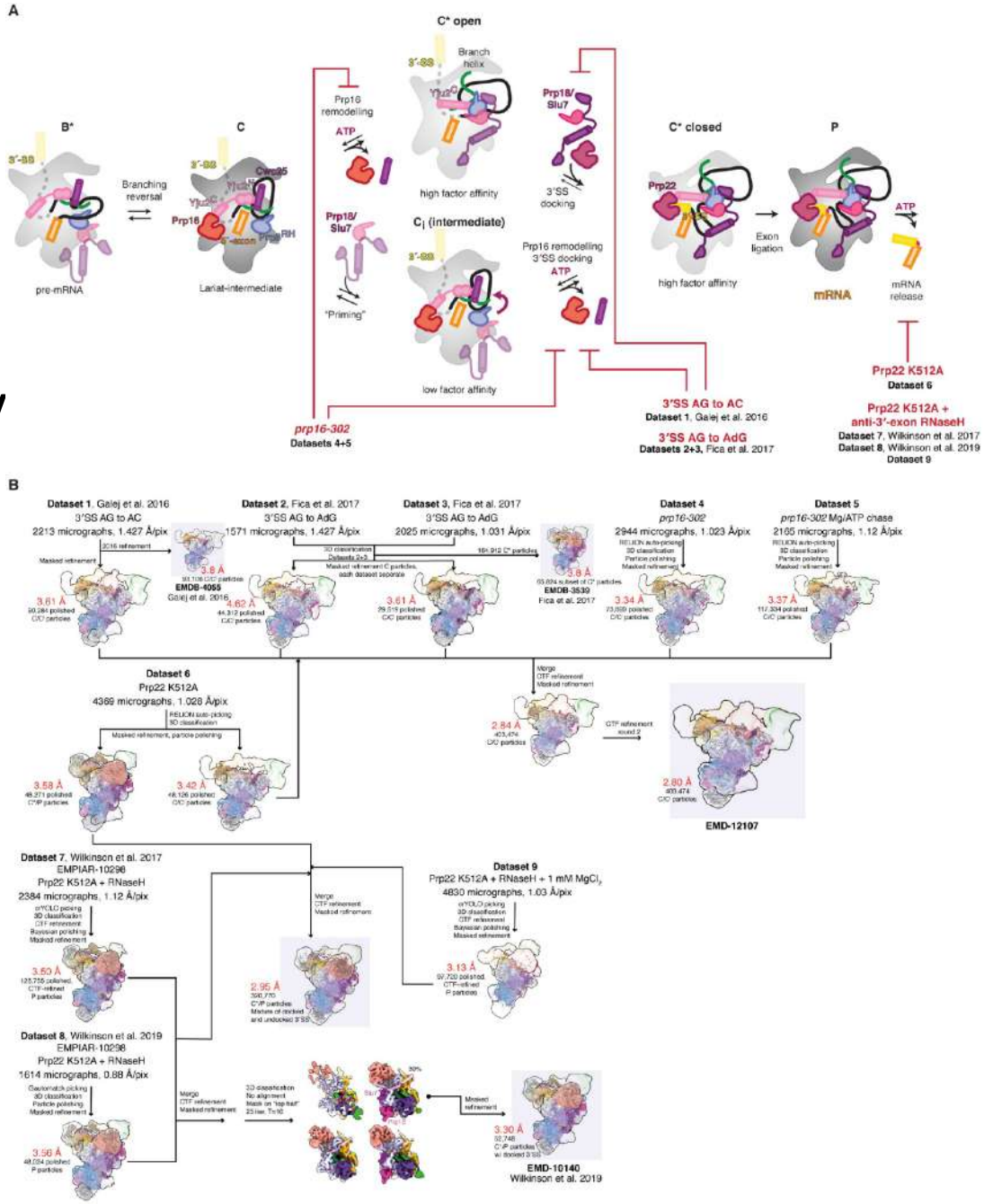
Compositional and conformational differences

- Human erythrocyte ankyrin-1 complex
- Multiple ab initio models generated
- 3D classification to identify compositional differences
- Signal subtraction and local refinement used to identify conformational differences



Vallese et al., 2022

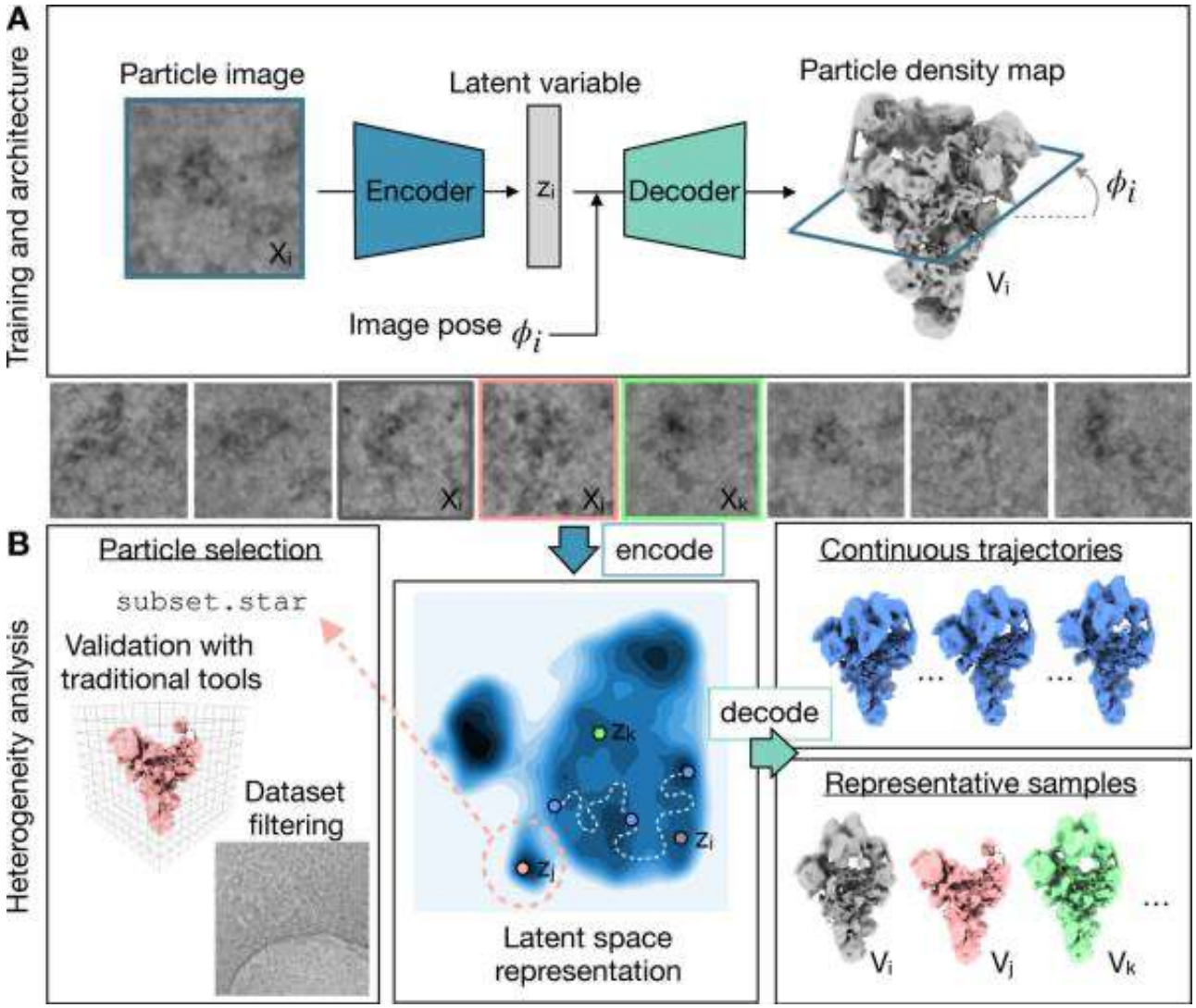
- Spliceosome
- Initial reconstruction is at 2.8 Å; however, lots of domains/proteins at periphery have poor density
- Signal subtraction coupled with focused classification and empirically determined tau2_fudge values (Relion) improve their resolutions
  - Try 1, 5, 10, 20, 50, 100, 200, 500, 1000 in parallel
- Relion



Wilkinson et al., 2021

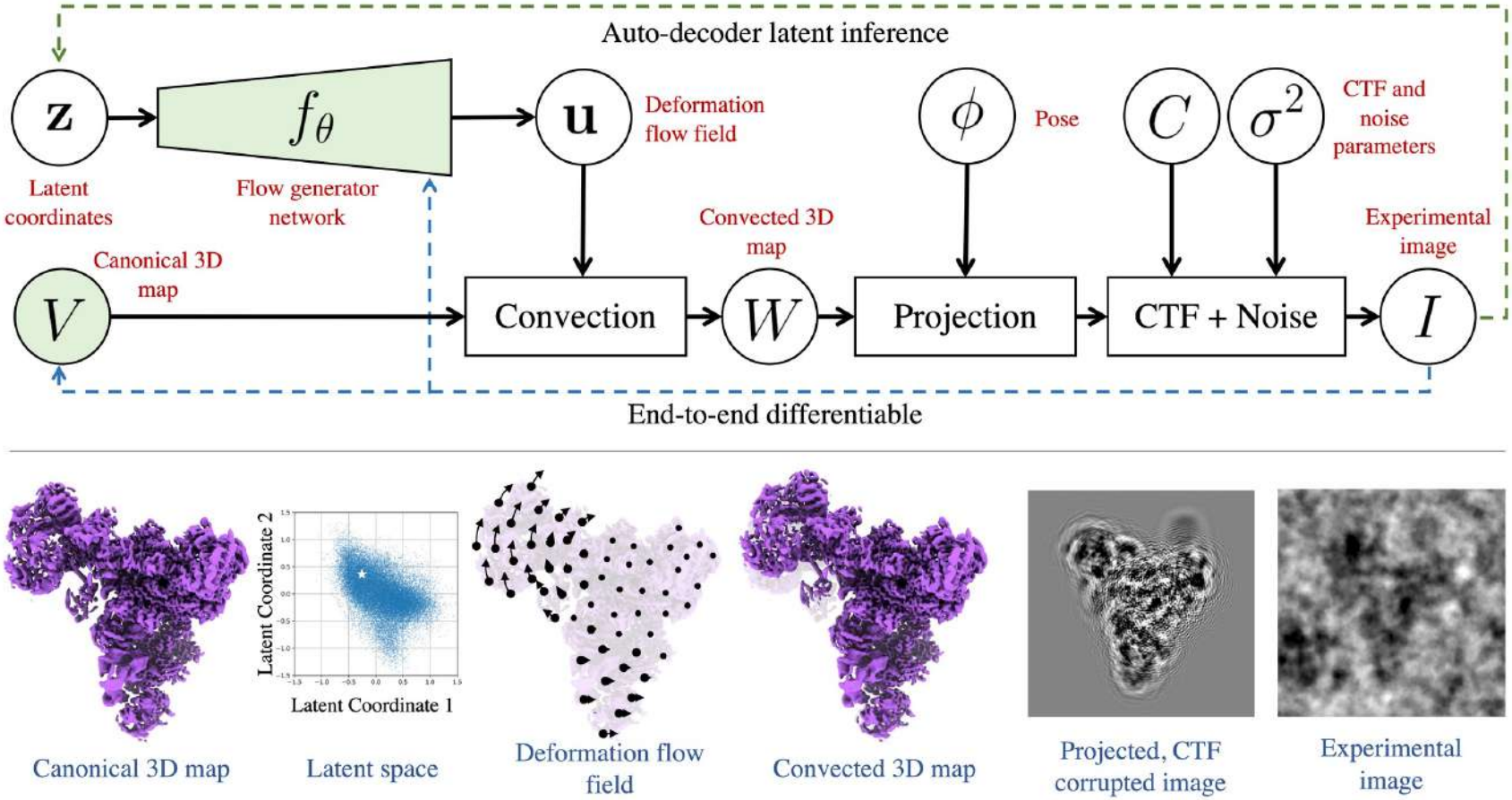
Continuous motion

- Spliceosome
- Uses deep generative model to identify data heterogeneity
- The latent space representation (contour map in bottom center) can be used to generate density maps
- Continuous trajectories can be generated for studying motion
- CryoDRGN



Zhang et al., 2021

3D Flexible refinement



- Uses deep generative model for continuous heterogeneity
- The user defined nonrigid deformation flow field can be used to improve resolution of flexible regions
- CryoSPARC

Punjani and Fleet 2023